# 关联性动态加权的协同过滤推荐

王　剑，余青松

(华东师范大学 计算机科学与软件工程学院，上海 200333)

**摘　要：**利用传统的协同过滤（CF）算法进行推荐时，由于用户评分矩阵比较稀疏，直接得到的用户或者项目之间的相似度相对而言可信度就比较低。为了解决这个问题，在传统的协同过滤基础上，引入项目与项目之间的关联性，通过在项目的类别标签和二部图的方法之间构建动态权重因子来融合这两种关联，形成非对等关联性关系，并做出用户对项目的评分预测，从而解决评分矩阵过于稀疏的问题。研究结果表明，相比于传统方法中使用对等相似度关系以及固定权值的方法，通过动态权重融合关联性形成非对等的关系的方法，更贴合生活实际，并且有更好的推荐效果。

**关键词：**协同过滤；评分矩阵；稀疏；动态权重因子；非对等关联性

**中图分类号：**TP301.6　　doi: 10.3969/j.issn.1001-3695.2018.05.0306

# Relevance dynamic weighted collaborative filtering recommendation

Wang Jian, Yu Qingsong

(*College of Computer Science & Software Engineering*, *East China Normal University*, *Shanghai* 200333, *China*)

**Abstract:** When the traditional collaborative filtering algorithm is used for recommendation, the credibility based on similarity between users or items directly obtained is relatively low due to the sparsity of the user rating matrix. In order to solve this problem, this paper introduced the relevance between projects on the basis of traditional collaborative filtering. The association is established in a non- reciprocal condition by building a dynamic weighting factor between the project's category label and bipartite graph approach, and the user's rating of the project will be predicted. As a result, compared with the method of using the equivalent similarity and the fixed weight value in the traditional method, the method of using non-equivalent relationship formed by the dynamic weights is more in line with the reality of life and has a better recommendation effect.

**Key words:** collaborative filtering; rating matrix; sparse; dynamic weighting factor ; non-equivalent relationship



## 0　引言

随着信息技术的迅速发展，互联网上产生了大量的数据，这给用户检索信息带来了困难和挑战。为了解决这个问题，对已有的信息进行过滤，就产生了推荐算法。推荐算法通常不需要明确的搜索查询目标，它能根据用户的一些特性，对信息进行筛选，为用户个性化地推荐理论上适合的项目。运用这样的算法进行信息过滤的系统模型通常称为推荐系统。推荐系统的历史可以追溯到 1979 年，与认知科学相关[1]，现在被广泛应用于各个领域，尤其是电子商务领域，其重要性尤为显著。

但是，在推荐系统中通常会由于评分矩阵过于稀疏而导致项目之间相似度计算的不准确，因此考虑对评分矩阵进行预测填充，利用填充后的评分矩阵对用户进行推荐，以提高推荐的准确性。

## 1　相关研究

应用于推荐系统的方法有很多。基于内容的推荐[2]结合用户的年龄，性别等个人信息或者项目的一些属性标签为用户进行推荐；基于协同过滤的方法[3,4]不考虑用户或者项目的属性，而是利用用户的行为信息进行推荐；基于标签的方法[5]利用用户对项目所做的标签的相关性进行推荐；还有利用时间、地点等上下文信息进行的推荐[6]等。为了对推荐系统的性能作出改善，也涌现出了很多改进的模型。例如 Zhou 等人[7]提出了一种利用概率传播机制的二部图推荐算法，构建了一种同类项目之间单向的、不对等的权重关系，推荐的准确率和召回率效果显著。Nilashi 等人[8]通过 SVD（奇异值分解）通常能筛选出对结果影响最大的几个隐性特征的特点，对稀疏矩阵进行降维，并用 ontology（本体）技术来提高推荐的准确性。Fan 等人[9]通过 KNN 算法寻找用户的最近邻，根据最近邻的评分情况对稀疏矩阵中的空白值进行预测填充，以提高协同过滤的准确性。

**收稿日期：**2018-05-25；**修回日期：**2018-07-06

**作者简介：**王剑（1993-），男，江苏高邮人，硕士研究生，主要研究方向为推荐系统、数据挖掘等（wangjian_csh@163.com）；余青松（1965-），男，高级工程师，硕士研究生，主要研究方向为计算机应用、Web 应用技术.

上述方法大多根据用户行为数据进行建模，整体效果较好，但有一定的局限性，因此本文提出一种动态融合类别和二部图思想的方法（RDWCF），综合内容和行为数据，并进行动态加权，以取得更好的推荐效果。

## 2　关联性动态加权的评分预测 CF

### 2.1　协同过滤

协同过滤是目前应用比较广泛，且比较成功的推荐技术之一。它的基本思想是根据用户或者项目的最近邻居的评分情况，将项目推荐给有相似需求的用户，即通过相似度度量方法获取到用户或者项目的最近邻居集，根据最近邻居集跟用户未知项目的关系情况，对用户与未知项目的关系作出预测。用户和项目之间的关系通常用评分矩阵来进行表示，如表 1 所示。

表 1　用户—项目评分矩阵

| | $I_1$ | $I_2$ | … | $I_n$ |
|---|---|---|---|---|
| $U_1$ | $R_{11}$ | ? | … | $R_{1n}$ |
| $U_2$ | ? | ? | … | $R_{2n}$ |
| … | … | … | … | … |
| $U_m$ | ? | $R_{m2}$ | … | ? |

其中，$U_m$ 表示第 m 个用户，$I_n$ 表示第 n 个项目，$R_{mn}$ 表示用户 $U_m$ 对项目 $I_n$ 的评分，“？”表示用户和项目尚未产生关联，是需要进行预测的值。

协同过滤在推荐系统中的应用是存在着各种各样的问题的[12]。随着推荐系统应用方业务规模的扩大，用户数量和项目数量均呈指数级增长，因而导致传统的协同过滤算法的性能越来越差，主要原因在于数据的稀疏性。在很多的推荐系统中，用户量和项目量都非常大，但是与每个用户产生关系的项目数量却是有限的，因此产生的用户-项目评估矩阵非常稀疏，这些评价数据大约只占总数的 1%~2%。这必然会导致对用户之间或者项目之间相似度计算的不准确性，如两个用户评价数目都相对稀少的情况下，他们的共同评价项并不能很好的反映他们的兴趣，但是二者却有很高的相似度。传统的协同过滤推荐算法还存在冷启动的问题，即对新用户和新物品由于数据的缺失，不能进行很好的推荐。

### 2.2　关联性

相似度通常是对称的，即 A 和 B 的相似度为 S，那 B 和 A 的相似度也是 S。这里引入关联性的概念来表示这种不对等的关系。其实从经验中我们可以也发现，关系总是不对等的，从不同的角度来衡量，得到的关系权重也是不一样的，即 A 和 B 之间的关系是不对等的，因此引入不对等的关联性概念来表示两个项目之间的关系显得更为合理。

#### 2.2.1 基于类别标签的关联性

往往事物都会有关于其类别标签的属性，一部电影它的类型可能是喜剧，惊悚或者爱情等等，一本书的类型也会有类似的类别标签，如表 2 所示。用户往往会对其中某种或多种类型的项目有一定的偏好。因为这样的关系，就可以基于项目的类型标签建立项目之间的联系，在一定程度上为用户推荐类型相似的项目[10,11]。

表 2　项目—类型矩阵

| | $G_1$ | $G_2$ | … | $G_k$ |
|---|---|---|---|---|
| $I_1$ | 1 | 0 | … | 0 |
| $I_2$ | 1 | 0 | … | 1 |
| … | … | … | … | … |
| $I_n$ | 0 | 1 | … | 0 |

表中 $I_n$ 表示第 n 个项目，$G_k$ 表示第 k 个类别，表中的数值 1 表示项目 I 属于这个类别，0 则表示不属于这个类别。

这里可以利用 Jaccard 公式建立起两个项目 $I_i$ 和 $I_j$ 的关联关系：

$$\mathrm{rel}_g(I_i,I_j)=\frac{|S(I_i)\cap S(I_j)|}{|S(I_i)\cup S(I_j)|} \tag{1}$$

其中：$S(I_i)$表示 $I_i$ 所对应的类别标签的集合。这是通过项目之间内在的属性建立的关系，因而是一种对等的关系。

#### 2.2.2 基于二部图的关联性

在推荐系统中，用户与用户，项目与项目以及用户与项目之间都存在着一定的关系，自然地可以想到用图来表示这种关系。把用户和项目都看作节点，他们之间的关系用线来表示，并在线上加上权重，用来表示用户对项目的评分情况。这样，我们就可以构建一个二部图来表示用户及项目之间的关系。二部图的基本模型如图 1 所示。

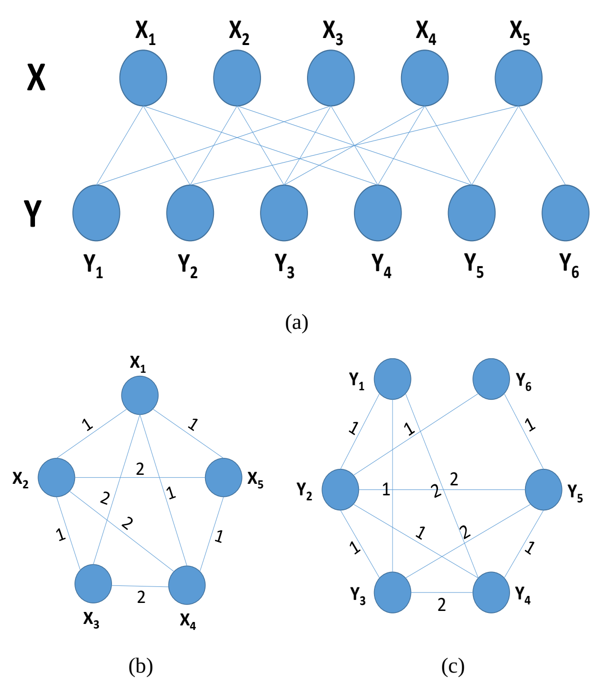



图 1　二部图算法的基本原理

图 1(a)中 X 表示用户节点，Y 表示项目节点。通过这样的图关系，我们就可以建立起用户与用户关系如图 1(b)，或者项目与项目之间的关系如图 1(c),连线当中的值表示它们共同的关系节点数。

二部图在推荐系统中的应用是利用了一种概率传播的机制。如图 2 所示。对于初始资源图 2(a)中的 A，B，C，以等概

率的方式向下传播，即将资源以等概率的方式平均地分配给下层的节点，如图 2(b)，同样的，下层节点再将资源以相同的方式回传给上层节点，如图 2(c)。其中"等概率"是指资源都是以平均分配的方式传递给与其有关联的节点的。同样也可以是"非等概率"的，即在推荐系统中，可以根据用户和项目之间的评分值来决定下层节点所占资源的比重。通过这样的传播方式，就将节点的资源混合了起来，每个资源都混合有其他节点的资源，因此就能表示出同类节点之间的关系。并且由于每种资源的混合程度不一样，这种关系也就成了一种不对等的关系，正因为这种不对等性，也更有利于我们进行个性化推荐。

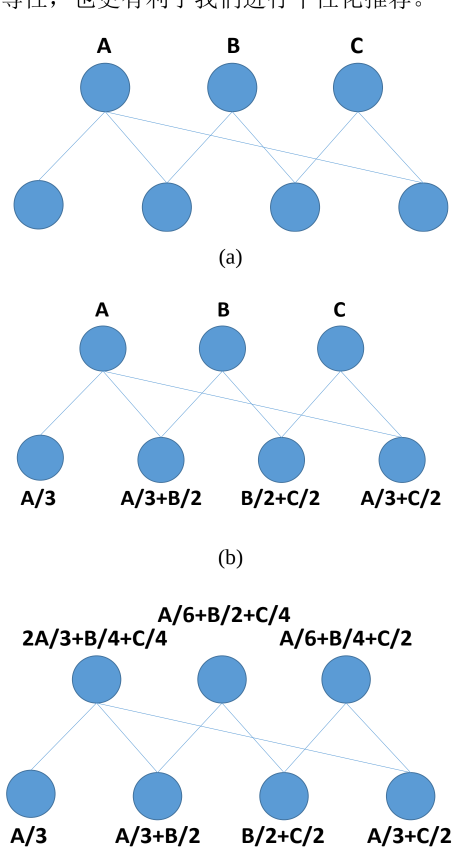



图 2　二部图算法的概率传播

**2.2.3** 关联性动态加权

前面提到了两种关联性计算方法，一种是基于内容的，一种是基于用户行为的，考虑将两种方式进行融合。

引入权重因子 $\lambda_1$ 和 $\lambda_2$，$\lambda_1$ 和 $\lambda_2$ 是关于关联性的动态权重因子，其通过关联性之间的关系动态构建两种关联性度量方法的权重比例，从而使得结果更加准确和稳定。动态权重因子 $\lambda_1$ 和 $\lambda_2$ 主要考虑不同的关联性度量方式会给用户预测评分带来一定的影响，从而动态地权衡不同关联性计算方法的比重，使预测结果更倾向于关联性强的一方。其公式为

$$\lambda_1 = \frac{rel_b^2(I_i, I_j)}{rel_g^2(I_i, I_j) + rel_b^2(I_i, I_j)} \tag{2}$$

$$\lambda_2 = \frac{rel_g^2(I_i, I_j)}{rel_g^2(I_i, I_j) + rel_b^2(I_i, I_j)} \tag{3}$$

其中：$\lambda_1+\lambda_2 = 1$，$rel_b(I_i, I_j)$是基于二部图概率传播的关联性，$rel_g(I_i, I_j)$是基于类别标签的关联性，$\lambda_1, \lambda_2$ 就是动态权重因子，通过其将两种关联性公式进行融合：

$$\text{rel}(I_i, I_j) = \lambda_1 * rel_b^2(I_i, I_j) + \lambda_2 * rel_g^2(I_i, I_j) \tag{4}$$

显然，这也是一种不对等的关联性关系。对于完全没有产生用户行为的项目，基于类别标签的关联度将占绝对的比重，也就是可以通过项目自身属性建立关联性，可以在一定程度上缓解推荐系统新项目上的冷启动问题。

### 2.3　评分预测

根据关联性计算方法，可以得到项目 I 的关联项目列表 M={$I_1$, $I_2$, …, $I_n$}，I∉M，并且关联性由高到低降序排列，即 $I_1$ 是跟 I 关联度最高的项目，$I_2$ 次之，依次类推。由此可得，用户 u 对未评分项目 I 的预测评分公式为

$$P_{u,I} = \bar{R}_I + \frac{\sum_{I_n \in M} rel(I, I_n) * (R_{u,I_n} - \bar{R}_{I_n})}{\sum_{I_n \in M} |rel(I, I_n)|} \tag{5}$$

其中：$\bar{R}_I$，$\bar{R}_{I_n}$ 是项目 I 和项目 $I_n$ 已获评分的平均得分，rel(I, $I_n$)是 I 和 $I_n$ 的融合关联性程度，$R_{u,I_n}$ 是用户 u 对 $I_n$ 的评分。

结合式（4）（5）可以看出，λ 和邻居数目都是实验结果的自变量，都对实验结果有一定的影响。而从 λ 的计算公式（2）（3）可以看出，λ 的结果采用的是整体的数据来进行计算，近邻数的选择并不对其产生影响，因此他们之间是两个独立的自变量。对于集合 M，将在实验中选择合适的元素数，避免加大程序的运算量，优化程序效率。

### 2.4　算法步骤

a)输入数据，建立用户-项目评分矩阵以及项目-类别矩阵。

b)利用项目-类别矩阵计算项目之间的内容关联性 $rel_g(I_i, I_j)$，式(1)。

c)利用用户-评分矩阵结合二部图方法计算项目之间的行为关联性 $rel_b(I_i, I_j)$。

d)计算动态权重因子 $\lambda_1$ 和 $\lambda_2$，公式(2),(3)。

e)利用式(4)对两种关联性进行融合，得到融合关联性 rel($I_i$, $I_j$)。

f)预测用户对未评分项的评分 $P_{u,I}$，公式(5)。

g)通过测试集对结果进行评估。

## 3　实验结果及分析

### 3.1　实验数据

为了验证本文提出的算法的推荐效果，采用了 MovieLens 电影评分数据集进行实验验证。该数据集包含用户对电影的评分数据，分值在 1～5 之间，分值越高表示用户对该电影越喜欢。除此之外，数据集中还包含电影的类型数据，包括 Action，Adventure，Animation，Children's，Comedy，Crime，Documentary，Drama，Fantasy，Film-Noir，Horror，Musical，Mystery，Romance，Sci-Fi，Thriller，War，Western 等 18 个类型，每部电影可以包含一个或多个类型。

本文选用的是 Movielens-100k 的数据集，该数据集中包含 943 个用户和 1682 部电影，评分数量大约有 10 万个，稀疏度大约为 93.70%，如表 3 所示。

表 3　数据集信息

| 用户数量 | 电影数量 | 评分数量 | 稀疏度 |
|---|---|---|---|
| 943 | 1682 | 10 万 | 93.70% |

从表中数据可以得到，用户的评分数据十分稀疏，问题的解决很有必要。在实验中，随机选取 80%的数据作为训练集，则剩下的 20%作为测试集，并多次实验交叉验证，避免实验的偶然性，以提高模型的性能。

### 3.2　检验标准

MAE(平均绝对误差)反映的是预测值与实际值误差的平均情况的度量值，常用来衡量推荐系统中评分预测结果的效果。其公式为

$$MAE = \frac{\sum_{i=1}^{n}|p_i - q_i|}{n} \qquad (6)$$

其中：n 为算法预测的评分项目的数目，$p_i$ 为预测评分，$q_i$ 为用户的实际评分。MAE 值越小证明与用户评分越接近，即 MAE 值越小，算法效果越好。在实验中，为了验证结果的准确性，通常进行多次实验，最终结果为多次实验的均值。

### 3.3　实验结果

计算采用文献[11]中的非动态权重因子方法来融合关联性方法的效果，$\lambda_1$，$\lambda_2$ 的取值范围为[0,1]，并且 $\lambda_1+\lambda_2 = 1$，则我们设置的 $\lambda_1$ 取值从 0.1 到 1，相应的 $\lambda_2$ 的取值为 0.9 到 0，相邻项目集取的是用户所有产生过关系的项目集。效果如图 3 所示。

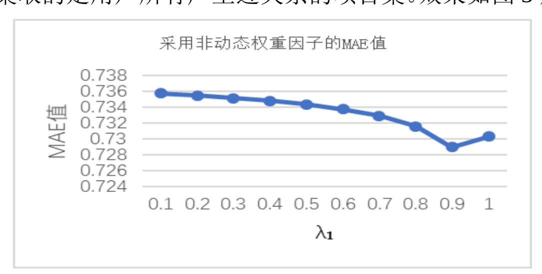



图 3　采用非动态权重因子方法的 MAE 图像

可以看到，当 $\lambda_1$ 取 0.9 时，采用非动态权重因子方法的 MAE 值达到最低。采用动态权重因子时，如图 4 所示，经过多次实验，实验的 MAE 值基本都是在采用非动态权重因子的方法的最低值之下，其均值也是远低于非动态权重因子的最低值的。

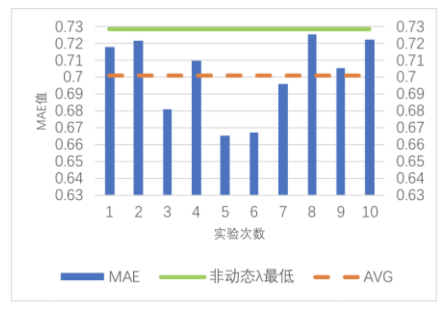



图 4　动态权重因子多次试验结果

其中，由于每次数据集划分的不同，所得出的结果也不尽相同，而柱状图表示的是采用动态权重因子多次实验得出的不同的 MAE 值，虚线表示它们的均值，实线表示采用非动态权重因子方法的最低值。

与传统的方法作对比，本文提出的方法也具有一定的优势，如图 5 所示。

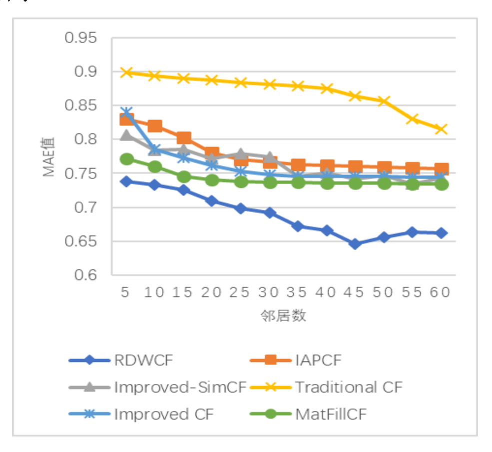



图 5　与其他算法的比较

分别与文献[13~16]中提出的方法以及传统的协同过滤方法在选择了不同近邻数的情况下做了比较，图中值为多次实验的均值，因此整体图像稍有波动，但整体效果良好。可以看到当近邻数为 45 左右的时候，本文提出的方法(RDWCF)的 MAE 值达到最低，变化减小并逐渐平稳。因此可以知道，动态权重因子以及关联性计算方法的引入，可以在一定程度上提高推荐系统的效果。

## 4　结束语

本文探讨了动态权重因子以及关联性对推荐系统的影响。首先介绍了推荐系统的相关研究，然后对本文用到的一些方法进行了系统性的介绍，最后从多个方面并多次实验进行比对。研究结果表明：二部图的关联性强度比标签的关联性强度要大，也就是用户行为的重要性要高于项目属性本身；使用动态权重因子比非动态权重因子效果好；动态权重因子结合关联性的方法比传统的融合方法更能提升推荐系统的性能。

## 参考文献：


[1] Rich E. User modeling via stereotypes [J]. Cognitive Science, 1979, 3 (4): 329-354.

[2] Pazzani M J, Billsus D. 10 content-based recommendation systems [M]. Berlin: Springer-Verlag, 2007: 325-341.

[3] Breese J S, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering [C]// Procof the 14th Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers Inc. 1998: 43-52.

[4] Zhang Baofu, Yuan Baoping. Improved collaborative filtering recommendation algorithm of similarity measure [C]// Proc of International Conference on Materials Science. New York: AIP Publishing

LLC, 2017: 109-132.

[5] Vig J, Sen S, Riedl J. Tagsplanations: explaining recommendations using tags [C]// Proc of International Conference on Intelligent User Interfaces. New York: ACM Press, 2009: 47-56.

[6] Adomavicius G, Tuzhilin A. Context-aware recommender systems [C]// Proc of ACM Conference on Recommender Systems. New York: ACM Press, 2008: 335-336.

[7] Zhou Tao, Ren Jie, Medo M, *et al.* Bipartite network projection and personal recommendation [J]. Physical Review E: Statistical Nonlinear & Soft Matter Physics, 2007, 76 (2): 046115.

[8] Nilashi M, Ibrahi O, Bagherifard K, *et al.* A recommender system based on collaborative filtering using ontology and dimensionality reduction techniques [J]. Expert Systems with Applications, 2017, 92.

[9] Fan Jiaqi, Pan Weimin, Jiang Lisi. An improved collaborative filtering algorithm combining content-based algorithm and user activity [C]// Proc of International Conference on Big Data and Smart Computing. Piscataway, NJ: IEEE Prsee, 2014: 88-91.

[10] 刘健，张琨，陈旋．基于标签和协同过滤的个性化推荐算法 [J]. 计算机与现代化，2016 (2): 62-65. (Liu Jian, Zhang Kun, Chen Xuan. Personalized recommendation algorithm based on tag and collaborative filtering [J]. Computer and Modernization, 2016 (2): 62-65. )

[11] 张景龙，黄梦醒，张雨，吴庆州．基于标签优化的协同过滤推荐算法 [J/OL]. 计算机应用研究 2018, 35 (10) . [2017-09-27]. http://www. arocmag. com/article/02-2018-10-023. html. (Zhang Jinglong, Huang Mengxing, Zhang Yu, Wu Qingzhou. Collaborative Filtering Recommendation Algorithm Based on Label Optimization [J/OL]. Application Research of Computers, 2018, 35 (10) . [2017-09-27]. )

[12] 孙小华．协同过滤系统的稀疏性与冷启动问题研究 [D]. 杭州：浙江大学，2005. (Sun Xiaohua. Research on sparsity and cold start of collaborative filtering system [D]. Hangzhou: Zhejiang University, 2005. )

[13] 王明佳，韩景倜．基于用户对项目属性偏好的协同过滤算法 [J]. 计算机工程与应用，2017, 53 (6): 106-110. (Wang Mingjia, Han Jingti. Collaborative filtering algorithm based on user preference for project attributes [J]. Computer Engineering and Applications, 2017, 53 (6): 106-110. )

[14] Wu Yueping, Zheng Jianguo. A collaborative filtering recommendation algorithm based on improved similarity measure method [C]// Proc of IEEE International Conference on Progress in Informatics and Computing. Piscataway, NJ: IEEE Press, 2011: 246-249.

[15] Liu Jianping, Wang Yong, Yan Fenghua. An improved collaborative filtering recommendation algorithm [C]// Proc of the 1st International Conference on Networking and Distributed Computing. Piscataway, NJ: IEEE Press, 2010: 204-208.

[16] 潘涛涛，文锋，刘勤让．基于矩阵填充和物品可预测性的协同过滤算法 [J]. 自动化学报，2017, 43 (9): 1597-1606. (Pan Taotao, Wen Feng, Liu Qinrang. Collaborative filtering algorithm based on matrix filling and item predictability [J]. Acta Automatica Sinica, 2017, 43 (9): 1597-1606. )